SEQUENTIAL PATTERNS AND TEMPORAL PATTERNS FOR TEXT MINING

By

Apirak Hoonlor

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

Approved by the
Examining Committee:

_____
Dr. Boleslaw K. Szymanski, Thesis Adviser

_____
Dr. Mark Goldberg, Member

_____
Dr. Mohammed J. Zaki, Member

_____
Dr. William A. Wallace, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2011
(For Graduation August 2011)

ABSTRACT

With the current growth rate of URLs, as a community, we are at the age of online information overload and for many other domains, such as Internet, web services, data analysis, and the like – they have been for quite sometimes. Text mining has been a key research topic for online information retrieval and information extraction. In this thesis, we studied two concepts in pattern extraction for text mining tasks: sequential pattern mining and bursty information.

In sequential pattern mining, our interests stemmed from a text mining problem of recognizing a group of authors communicating in a specific role within an Internet community. The challenge is to recognize possibly different roles of authors within a communication community based on each individual exchange in electronic communications. Depending on the exchange parties, the message can vary in length, contain different style of writing, and contain multiple topics, making the standard text mining approaches less efficient than in other applications. An example of such a problem is recognizing roles in a collection of emails from an organization in which middle level managers communicate both with superiors, subordinates and among themselves.

For this problem, we present *Recursive Data Mining* (RDM), a sequence pattern mining framework for text data. RDM allows certain degree of approximation in matching patterns – necessary to capture non-trivial features in text datasets. The framework minimizes the size of the combinatorial search space using statistically significant tests, such as information gain, mutual information, and minimum support. RDM recursively and hierarchically mines patterns at varying degrees of abstraction. From one abstraction level to another, the framework removes "noisy" tokens to allow long range patterns to be discovered at higher levels. We used a hybrid approach, in which the RDM discovered patterns are used as features, to build a classifier. We validated RDM framework on the role identification task using Enron email dataset. Specifically, we used RDM to categorize the senders of email content into their roles in Enron as CEO, Vice-President, Manager and Traders. The results

showed that a classifier that used the patterns discovered by Recursive Data Mining performs well in role identification.

Our interests in the concept of bursty information originated from the temporal nature of social roles of each individual. Over a life time, a person can be associated with variety of roles ranging from a leader of an arm group, a prisoner, a Nobel Prize winner to a president of a country. Such information is often embedded in, and can be extracted from, the temporal text patterns associated with an individual. Previously, the term frequency was the main quantifier for trend analysis and visualization. In recent years, due to their abilities to detecting changes in patterns, the concepts of bursty information have been used to extract temporal patterns from text streams. We proposed two complementary burstiness frameworks to extract temporal correlated patterns from text stream. The first framework is proposed for the extraction of bursty patterns in the bursty period of a given pattern. The second framework is proposed for the extraction of temporally correlated patterns at each time steps. We use these frameworks to analyze ACM dataset. Specifically, we used them to find out the following: (i) when certain research topics received high interests from Computer Science research communities, and during which time, what were the related topics often associated with them, (ii) for each topic, which other topics are temporally correlated with it at a given time.

As we studied the burstiness concepts, we realized that they can be applied to other text mining tasks. We proposed a *bursty distance measurement* for creating a distance matrix in text clustering task. We experimented with our framework on synthetic data, online news article data, and additional real-life datasets. The experiments showed a substantial improvement on event-related clustering on online news article data for our framework. Also, our framework generally performed better than other existing methods. In the future, we intend to embed the bursty information into our RDM framework and use it to trace the changes, or lack thereof, in the sequential patterns and hierarchical structures of text streams. We also want to see if we can incorporate more information into sequential patterns using the idea of surface text pattern.